Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Modeling individual-level long-range brain dynamics from short fMRI scans

Yixuan Yuan

yxyuan@ee.cuhk.edu.hk

The Chinese University of Hong Kong https://orcid.org/0000-0002-0853-6948

Yu Jiang

The Chinese University of Hong Kong

Cheng Wang

The Chinese University of Hong Kong

Zhihao Peng

The Chinese University of Hong Kong

Zhibin He

Northwestern Polytechnical University

Boyun Zheng

The Chinese University of Hong Kong

Yifei Sun

University of Sydney

Tuo Zhang

Northwestern Polytechnical University

Peter Yat Ming Woo

Prince of Wales Hospital

Robert D Sanders

University of Sydney

Fernando Calamante

University of Sydney

Jinglei Lv

University of Sydney

Article

Keywords: fMRI, Deep learning in neuroimaging, Zero-shot transferability, Large language models

Posted Date: April 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9217734/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Modeling individual-level long-range brain dynamics from short fMRI scans

Yu Jiang[1], Cheng Wang[1], Zhihao Peng[1], Zhibin He[2], Boyun Zheng[1], Yifei Sun[3], Tuo Zhang[2], Peter Yat Ming Woo[3], Robert D Sanders[4], Fernando Calamante[5], Jinglei Lv[4,5*], Yixuan Yuan[1*]

[1*]Department of Electronic Engineering, The Chinese University of Hong Kong, Ma Liu Shui, Shatin, 000000, Hong Kong, China.
[2]School of Automation, Northwestern Polytechnical University, West Youyi Road, Xi'an, 710072, Shaanxi, China.
[3]Department of Surgery, Prince of Wales Hospital, Yin Cheng Street, Hong Kong, 000000, Hong Kong, China.
[4]Central Clinical School, University of Sydney, NSW 2050, Australia.
[5]School of Biomedical Engineering, University of Sydney, NSW 2050, Australia.

*Corresponding author(s). E-mail(s): jinglei.lv@sydney.edu.au; yxyuan@ee.cuhk.edu.hk;
Contributing authors: yujiang@cuhk.edu.hk; chengwang@link.cuhk.edu.hk; zhihaopeng@cuhk.edu.hk; hezhibin@main.nwpu.edu.cn; byzheng@link.cuhk.edu.hk; yifei.sun@sydney.edu.au; tuozhang@nwpu.edu.cn; peterwoo@surgery.cuhk.edu.hk; robert.sanders@sydney.edu.au; fernando.calamante@sydney.edu.au;

Abstract

Functional magnetic resonance imaging (fMRI), which enables noninvasive measurement of Blood Oxygen Level Dependent (BOLD) brain dynamics, is a cornerstone for studying human brain function and connectivity. In real-world settings, however, fMRI acquisitions are often constrained by limited scanner time and subject compliance. These short scans frequently yield unstable functional connectivity estimates, limiting the reliability of brain-based markers (e.g., connectome) and reducing confidence in downstream tasks. Existing computational approaches only partially address this challenge, as they often struggle to achieve both reliable individual-level reconstruction and strong cross-cohort generalization from short observational windows. Here, we present BOLD-Cast, a computational framework for reconstructing long-range BOLD dynamics from short fMRI scans. BOLD-Cast jointly learns cohort-invariant and subject-specific functional topology to enable individualized reconstruction with explicit temporal anchoring. Trained on UK Biobank data ($n > 30000$), the framework preserves highest functional connectivity topology with a similarity score of 0.996, while supporting stable reconstruction over sequences up to 2 times the input length. Moreover, our approach demonstrates strong zero-shot generalization on 4 unseen independent datasets (HCP-YA, HCP-A, HCP-D, ABIDE), spanning adolescence to older adulthood and cohorts with neuropsychiatric conditions. We further show that reconstructed fMRI signals improve autism spectrum disorder classification accuracy by 7.2% compared with using short scans alone, and enhance cognition-related prediction performance. Together, these results establish a generalizable framework for improving brain-based inference from short-duration fMRI across heterogeneous cohorts.

Keywords: fMRI, Deep learning in neuroimaging, Zero-shot transferability, Large language models.

1 Introduction

According to statistics reported by the World Health Organization (WHO), more than one billion people worldwide suffer from mental health disorders [1]. Pinpointing the mechanisms under the symptoms of disorders requires reliable measurement of brain functional organization, which becomes an urgent priority for intelligent medicine. Functional magnetic resonance imaging (fMRI) has played a crucial role in supporting clinical decisions, e.g., detection of functional connectivity abnormalities in autism [2–5], support stroke recovery [6, 7], preoperative planning for epilepsy and brain tumors [8–10], and evaluation of drug efficacy and guiding personalized transcranial magnetic stimulation for depression [3, 11, 12]. These tasks rely on temporal accuracy and intra-subject/inter-scan stability. Studies have shown that for every minute longer scan time, intraclass consistency (ICC) improves by over 20%, significantly enhancing the confidence and test-retest reliability of functional connectivity (FC) [13, 14].

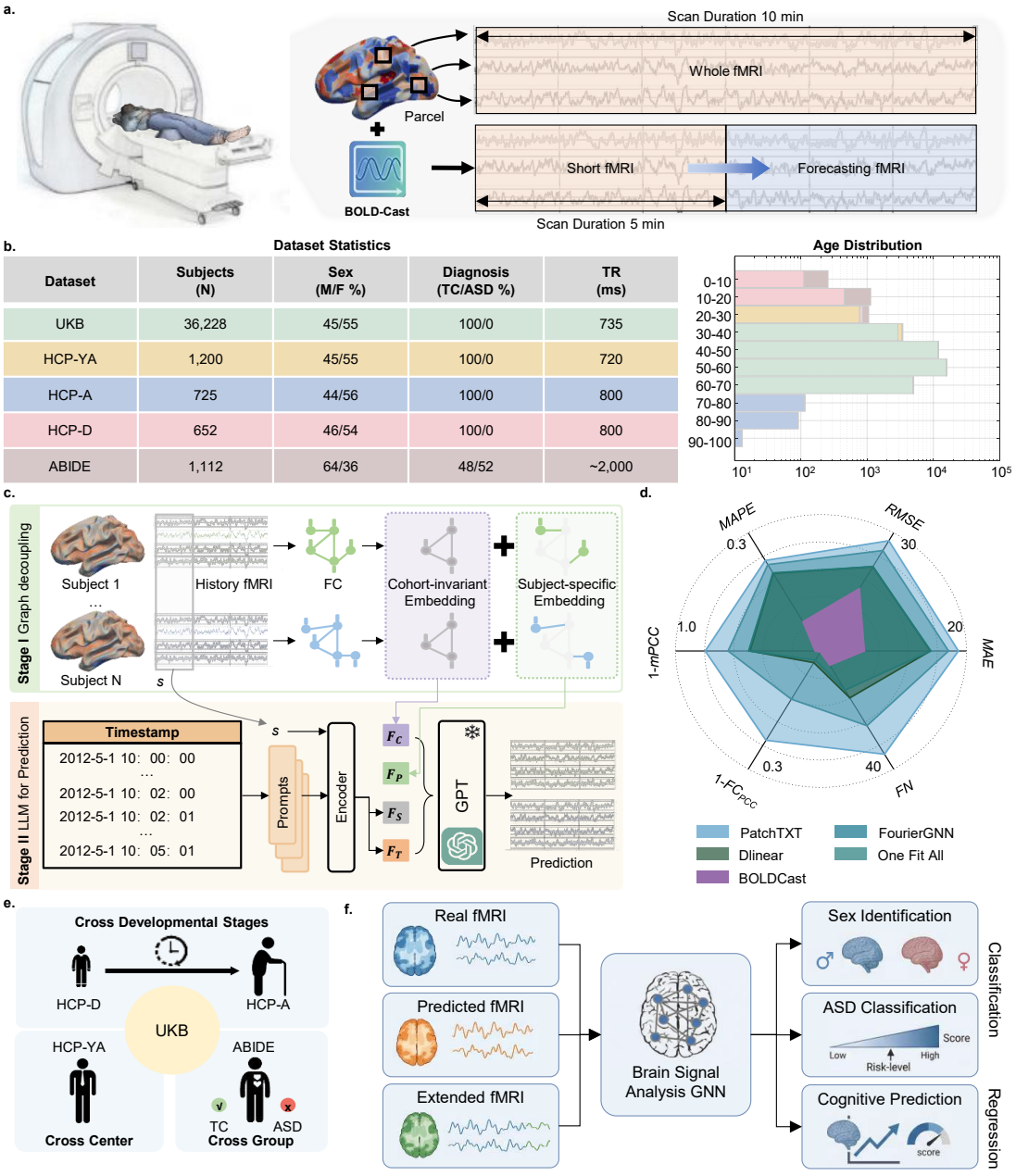
However, long high-quality scans remain rarely feasible in real-world clinical settings: children, the elderly, and patients with neuropsychiatric disorders often struggle with compliance [15–19]. Large-scale cohorts (e.g., the American Adolescent Brain Cognitive Development (ABCD) [20], the Autism Spectrum Disorders Multicenter Dataset (ABIDE) [21]) also compress scan times to under ten minutes to enhance feasibility. This unavoidable trade-off severely reduces fMRI's statistical power, leading to unstable connectivity estimates [15] and limited translational reliability and utility in clinical practice [16].

Several recent approaches have attempted to alleviate the limitations of short fMRI acquisition, which can be broadly classified into two categories. One is sequence acceleration technology aiming to improve sampling efficiency per unit time, such as GRAPPA [22], SENSE [23], multiband EPI [24], and Multiplexed-EPI [25]. Although effective in reducing acquisition duration, they often suffer from signal-to-noise ratio degradation and noise amplification (e.g., effects of the g-factor [26]), further deteriorating data stability and image quality. Another promising approach leverages deep learning models to predict long-range blood oxygen level-dependent (BOLD) dynamics from short periods [27–30], effectively extending the length of the fMRI time series without increasing the in-scanner time.

Despite recent advances, such prediction models face two key barriers to clinical translation. First, they often lack the capacity to generalize across heterogeneous populations. Clinical cohorts, such as elderly patients or adolescents with autism, exhibit functional connectivity patterns that deviate from healthy adults [31, 32]. Most models, while performing well on healthy adult data, fail to capture invariant brain networks shared across diverse health states and age groups [33–35], limiting their reliability in real-world diagnostic settings. Second, current models offer limited temporal flexibility. In practice, fMRI acquisition needs vary widely: rapid screening scenarios demand brief scans [36], while longitudinal assessments require extended observation [37]. Models restricted to fixed-length prediction windows are poorly suited to this spectrum of clinical demands, limiting utility across various diagnostic workflows [38].

To address these challenges, we introduce **BOLD-Cast** (short for BOLD fMRI forecast), a two-stage autoregressive framework that synthesizes biologically plausible, temporally stable fMRI time series from short scans (Fig. 1**c**). In Stage I, a graph-based disentanglement encoder decomposes the fMRI history of each subject into a cohort-invariant embedding and a subject-specific embedding, stabilizing shared functional topology while preserving individual variability. In Stage II, these embeddings are combined with sequence features and timestamp-based prompts and fed into an autoregressive model built on GPT [39], the open-source large language model (LLM), which forecasts future fMRI time points with explicit temporal anchoring. This design enables BOLD-Cast to generate long-range, temporally coherent dynamics from short scans in a cohort-agnostic manner. Our key innovations are twofold: (i) disentangling functional topology to support individualized forecasting, and (ii) repurposing the autoregressive capability of LLMs to predict biological brain dynamics.

BOLD-Cast is trained on the large neuroimaging dataset, UK Biobank (UKB) [40], enabling **high-fidelity forecasting**. We demonstrate that BOLD-Cast can reliably extend subsequent dynamics while preserving both temporal dynamics and spatial FC in Fig. 1**d**. We observe that even when predicting sequences twice the input length, the model maintains considerable performance, highlighting its **strong temporal scalability**. We further validate BOLD-Cast across four external datasets (HCP-Young Adult (HCP-YA) [41], HCP-Development (HCP-D) [42], HCP-Aging (HCP-A) [43], and ABIDE) spanning diverse acquisition protocols, age ranges, and



Dataset	Subjects (N)	Sex (M/F %)	Diagnosis (TC/ASD %)	TR (ms)
UKB	36,228	45/55	100/0	735
HCP-YA	1,200	45/55	100/0	720
HCP-A	725	44/56	100/0	800
HCP-D	652	46/54	100/0	800
ABIDE	1,112	64/36	48/52	~2,000

Fig. 1 Overview of BOLD-Cast framework for fMRI sequence extension and downstream task verification. **a**, Conceptual illustration of BOLD-Cast. The model operates on parcel time series and predicts the subsequent latent fMRI dynamics based on an initial short scan. **b**, Dataset statistics and demographic distribution. The benchmark includes five fMRI datasets (UKB, HCP-YA, HCP-A, HCP-D, ABIDE) with variation in age, sex, diagnostic category (TC: typical control, ASD: autism spectrum disorder), and scanning protocols (TR: repetition time). **c**, Two-stage architecture of BOLD-Cast. n Stage I, a graph-based embedding module separates the subject-invariant (common) and subject-specific (individual) components from the observed history fMRI. In Stage II, an LLM (frozen during training) is prompted with a sequence of learned embeddings concatenated with timestamp encodings, enabling autoregressive prediction of future latent fMRI representations with temporal awareness. The predicted embeddings are decoded into extended fMRI time series. **d**, Quantitative comparison using six metrics. BOLD-Cast achieves consistently lower error (MAE, $RMSE$, $MAPE$, FN) and higher correlation ($mPCC$, FC_{PCC}) than existing baselines on UKB. **e**, Generalization scenarios evaluated in this study. BOLD-Cast is tested on five datasets across imaging centers (e.g., UKB ↔ HCP-YA, UKB ↔ ABIDE), age stages (e.g., HCP-D → HCP-YA → HCP-A), and diagnostic groups (e.g., TC ↔ ASD). **f**, Overview of downstream tasks supported by the generated fMRI, including cognitive score prediction (regression), ASD classification and sex identification (classification).

clinical conditions. We consistently find that the BOLD-Cast learned population-invariant representations of FC that align closely with canonical resting-state networks, thereby supporting its **biological plausibility**, which in turn underpins its **cross-cohort generalizability**. We also

assess the **practical utility** of BOLD-Cast in downstream inference. By extending short scans in ABIDE, we observe improved accuracy and $F1$ scores for autism spectrum disorder (ASD) support-inference and sex identification, consistent with evidence that longer fMRI durations improve representational reliability [15]. We conclude that BOLD-Cast offers a robust, broadly applicable, and clinically compatible framework for scalable fMRI forecasting.

2 Results

Resting-state fMRI data from 26,000 UKB participants were employed to train BOLD-Cast, 9,228 were used to test model performance and 1000 for validation. BOLD signals were parcellated using the Craddock CC200 (CC200) atlas [44] and averaged to represent regional activity. An 81-TR (1 minute) sliding window was used to generate samples for training the model, and the model was trained to forecast the subsequent 81 TRs of BOLD activity using meansquared error (MSE) as the training loss. To comprehensively assess the generalizability and clinical adaptability of BOLD-Cast, we established a large-scale evaluation framework spanning multiple cohorts. On the population level, we established a large-scale evaluation framework spanning multiple cohorts, including HCP-YA (n=1200), HCP-D (n=652), HCP-A (n=725), and ABIDE (n=1,112), covering diverse age groups, acquisition protocols, and clinical conditions. As downstream tasks, we also evaluated its ability to support cognitive score prediction, ASD classification, and sex identification.

To ensure fair and comparable evaluation across tasks and cohorts, all experiments adopted a unified benchmarking benchmarking protocol including parcellation with CC200 atlas, a sliding-window forecasting setup, and optimization objective of minimizing MSE. All baseline models were retrained on the UKB dataset using the same training-testing split, thereby eliminating confounding effects of data source or sample size. This unified benchmarking protocol allows us to systematically quantify model performance across five key dimensions: predictive accuracy, temporal scalability, cross-cohort generalization, neuroscientific consistency, and clinical utility. The following sections are structured according to these five dimensions and demonstrate how BOLD-Cast consistently achieves the state-of-the-art (SOTA) performance across diverse medical neuroimaging scenarios.

2.1 BOLD-Cast Delivers High-Fidelity fMRI Forecasting

To validate BOLD-Cast's forecasting performance, we conducted a systematic comparison on the UKB dataset against 11 representative models spanning four major time-series paradigms:

- **Graph neural networks**: MSGNet [45], ForecastGrapher [46] and FourierGNN [47];
- **Transformers**: iTransformer [48], PatchTST [49] and BrainTransformer [28], a task-specific model tailored for fMRI prediction;
- **Linear models**: TSMixer [50], DLinear [51] and LightTS [52];
- **Large language model–inspired models**: SimMTM [53] and One Fit All [54].

These models were selected based on their proven success in large-scale time-series domains—such as power systems and transportation networks, which share structural analogies with regional fMRI time series. Model-specific hyperparameters are detailed in Supplementary Tables S9–S15. We assessed performance across two clinically salient dimensions: (i) temporal fidelity and (ii) preservation of functional connectivity.

In the setting of 81-TR prediction with 81-TR, BOLD-Cast achieved the best performance in predicting temporal dynamics measured by four standard metrics ($MAE = 5.96$, $RMSE = 15.80$, $MAPE = 0.074$, $mPCC = 0.96$, which is the average of Pearson Correlation Coefficients (PCC) calculated across different parcels), accurately recovering both the global trend and local fluctuations of BOLD trajectories (Fig. 2**a-b**). Among all competing baselines, One Fit All is the strongest competitor (ranked second overall after BOLD-Cast). Therefore, for brevity, subsequent qualitative and error-profile comparisons focus on this second-best model. Compared with One Fit All, BOLD-Cast significantly reduced phase shifts and amplitude deviation ($p < 0.01$), and exhibited more stable errors over time (Fig. 2**c**). The error distribution (Fig. 2**d**) further revealed a sharper concentration around zero and substantially suppressed long-tail deviations, indicating its superior capability in capturing subtle neural dynamics.

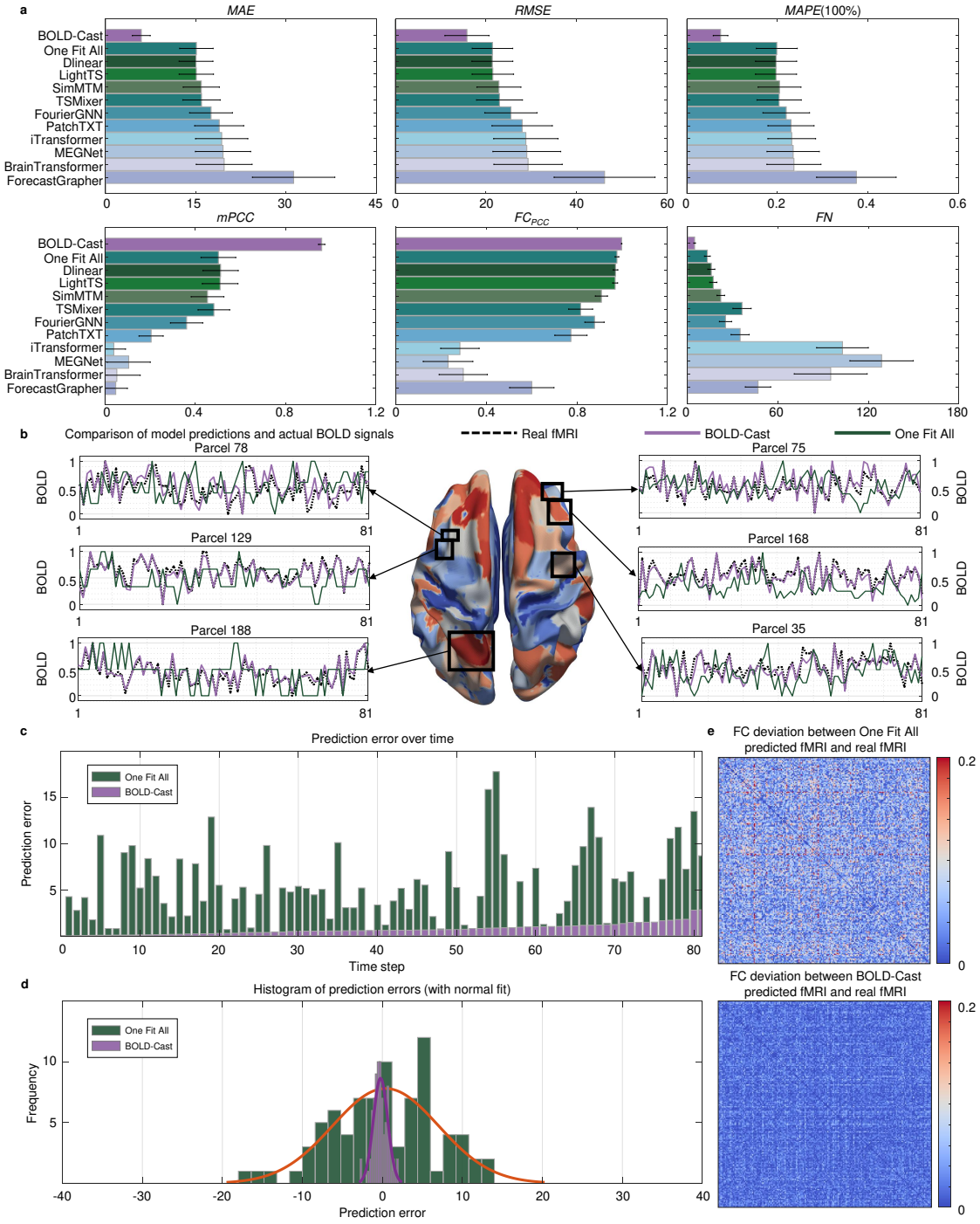


Fig. 2 Benchmarking BOLD-Cast on UKBfMRI dataset. **a**, Quantitative comparison with baseline forecasting models across six metrics. BOLD-Cast (purple) consistently outperforms One Fit All (green) and other baselines, achieving lower error and better FC alignment, with the smallest error bars observed across all metrics. **b**, Representative BOLD signal trajectories from six randomly selected brain regions from CC200 atlas. Each network is visualized by averaging the signals of its constituent brain regions. The brain regions are represented by scatter plots. The figure overlays the actual values (black), BOLD-Cast predicted values (purple), and the One Fit All model (green). Shaded bands represent inter-brain region differences. The BOLD-Cast model exhibits lower bias and better temporal alignment. **c**, Step-wise absolute prediction errors across forecast time. Compared to One Fit All, BOLD-Cast yields consistently smaller and more stable errors across the horizon. **d**, Histogram of prediction errors, aggregated across all samples. Errors from BOLD-Cast concentrate more tightly around zero. **e**, Heatmaps of FC deviation from ground-truth FC. BOLD-Cast shows substantially reduced FC differences compared to One Fit All, confirming superior preservation of brain connectomes.

BOLD-Cast also demonstrated excellent preservation of inter-regional connectivity. FC maps derived from its predicted fMRI showed higher similarity to ground-truth, as measured by the

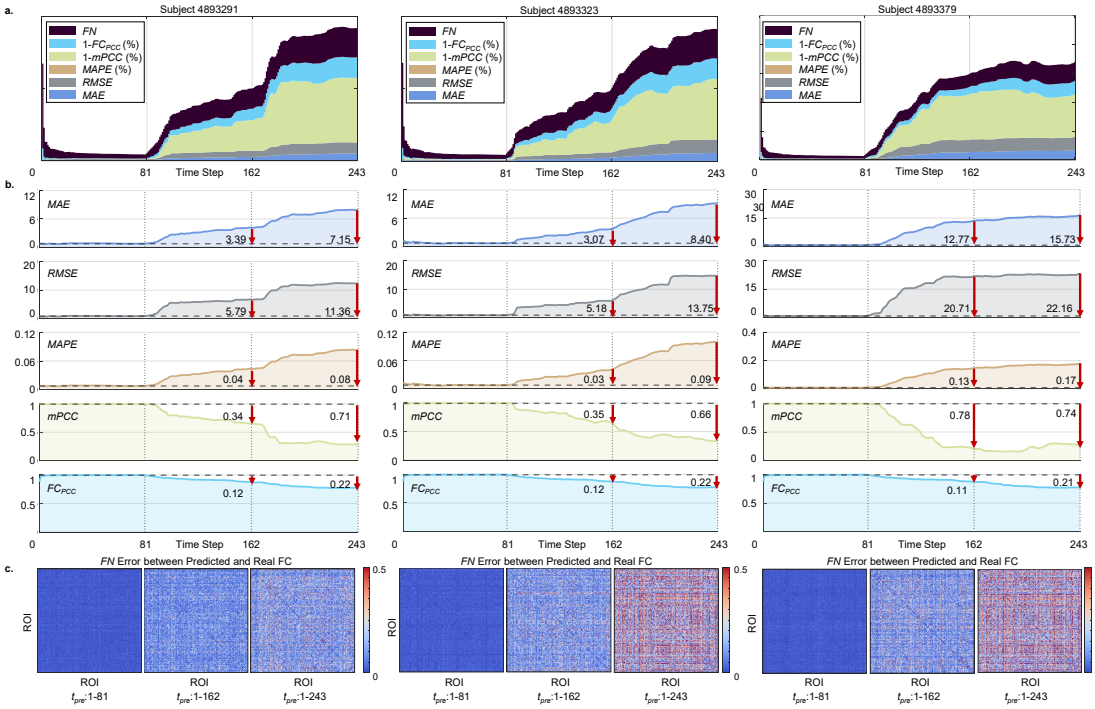


Fig. 3 Scalability of BOLD-Cast under extended prediction horizons. **a**, Stacked prediction errors across three representative subjects (from UKB) as forecast horizons extend from 81 to 243 time steps (3 training window). Each stacked bar represents the accumulated sum of multiple error metrics, with colored segments in order of MAE (purple, bottom), $RMSE$ (green), $1-mPCC$ (yellow), $1-FC_{PCC}$ (blue), $MAPE$ (orange), and FN (black, top), measured as the Frobenius norm between predicted and real FC matrices. Most metrics increase gradually, but FC connectome preservation (FC_{PCC} and $mPCC$) remains relatively stable until 160 steps. **b**, Temporal evolution of error metrics for each subject. The grey dashed line denotes the performance of BOLD-Cast at the training horizon (81 steps), serving as a reference for long-range prediction stability. Despite expected increases in MAE, $RMSE$, and $MAPE$ over time, structural metrics ($mPCC$ and FC_{PCC}) decline more slowly, indicating preserved network-level dynamics across longer horizons. **c**, FC deviation maps between predicted and real fMRI at time windows [1–81], [1–162], and [1–243]. Deviations remain minimal at 81 steps and increase with horizon length, but overall connectome remains preserved, demonstrating BOLD-Cast's ability to sustain coherent connectivity patterns even under extrapolated predictions.

highest similarity ($FC_{PCC} = 0.996$) and the lowest Frobenius norm ($FN = 0.028$) among all baselines. Heatmaps of the difference matrices (Fig. 2**f**) highlighted that BOLD-Cast more faithfully reconstructed brain-wide coupling patterns, with dominant blue (low error) regions and fewer red (high error) regions compared to One Fit All, underscoring its ability to preserve neurotopological consistency. The corresponding numerical results and visualizations of the results of all other compared models are provided in Table S1 and Figures S1–S3 in the Supplementary Material. Furthermore, the test results of BOLD-Cast on the HCP-YA, HCP-D, HCP-A and ABIDE dataset are provided in Table S2 in the Supplementary Material.

2.2 BOLD-Cast Scales Robustly to Longer Temporal Predictions

After confirming BOLD-Cast's high-fidelity performance on short-term term (1 minute) forecasting, we next assessed its robustness over longer prediction horizons, i.e., predict twice or triple of the input sequence, a key requirement for clinical scenarios with varying scan durations and temporal availability. To this end, while maintaining an 81-step observation input, we fixed the observation window at 81 steps and progressively increased the prediction horizon to 162 (2×) and 243 (3×) steps. This setup simulates realistic long-range fMRI forecasting, where future predictions are iteratively generated based on prior outputs. The corresponding numerical results and a more refined visualization of the model predictions across more parcels from steps 1 to 243 are shown in Supplementary Material Table S3.

Overall, BOLD-Cast exhibited stable performance up to twice the input length, with minimal degradation in accuracy (Fig. 3**a**). At the 2× horizon, error metrics (such as MAE) increased

gradually, while FC similarity metrics (such as FC_{PCC} and FN) remain high (Fig. 3**b**). Visualizations of FC difference (Fig. 3**c**) further revealed that although spatial distortions accumulate beyond 2×, the model largely preserves functional topography within this extended window. We attribute this robustness to BOLD-Cast's temporal conditions and autoregressive generation mechanism, enabling dynamic extension without compromising temporal smoothness or network-level consistency. This capability offers practical value for adapting to scenarios such as different sampling periods, enhancing BOLD-Cast's applicability in real-world clinical neuroimaging tasks.

2.3 BOLD-Cast Maintains Robust Generalization Across Diverse Domains

Following the validation of BOLD-Cast's prediction accuracy and scalability, we further explored its generalization capabilities across unseen neuroimaging cohorts, spanning diverse acquisition protocols, developmental stages, and clinical conditions. We assessed BOLD-Cast in a strict zero-shot setting: the model was trained solely on UKB without any fine-tuning or domain adaptation, and directly tested on four independent datasets: HCP-YA, HCP-D, HCP-A and ABIDE. To ensure comparability across datasets with different TRs, the prediction horizon was matched by real time (≈ 2 minutes) rather than step count. This corresponds to 162 steps on UKB, 166 on HCP-YA, 150 on HCP-A, 170 on HCP-D, and 60 on ABIDE. As shown in Fig. 4**a**, BOLD-Cast consistently outperformed 11 representative prediction models in both time-series precision and FC preservation across all settings (see Supplementary Table S4-S7 for numerical results).

Generalization Across Acquisition Protocols First, we examined the model's transfer performance across different acquisition conditions. The HCP-YA dataset is acquired with different scanners and protocols from UKB Despite hardware-induced domain shift, BOLD-Cast achieved the lowest prediction error ($MAE = 7.60$, $MAPE = 0.10$) and the highest correlation consistency ($mPCC = 0.96$, $FC_{PCC} = 0.99$) among all models (Fig. 4**a**). These results suggest that the model effectively captures stable neural dynamics shared across different acquisition environments. In contrast, other methods suffer from severe amplitude distortion and lower network consistency, highlighting BOLD-Cast's advantage in cross-center adaptability. The compact subject-wise error distribution (Fig. 4**b**) further supports its robustness to scanner variation.

Beyond HCP, we further evaluated cross-center generalization on the ABIDE dataset, which represents an even more challenging transfer scenario due to pronounced differences in acquisition protocols, scanner vendors, and site-specific preprocessing. As shown in Fig. 4**a** and Table S7 in Supplementary, BOLD-Cast remains the only method that yields stable and meaningful predictions when directly transferred from UKB to ABIDE, achieving substantially lower errors and higher functional connectivity consistency than all competing approaches. In contrast, most baseline models either exhibit severe performance degradation or fail to produce valid outputs on ABIDE, resulting in undefined (NaN) metrics. This failure likely reflects the sensitivity of these methods to protocol-specific signal scaling and temporal distortions, which are exacerbated under strong domain shifts. These results highlight BOLD-Cast's unique ability to generalize across both moderate (UKB→HCP) and extreme (UKB→ABIDE) acquisition mismatches, underscoring its suitability for real-world, multi-center neuroimaging applications.

Generalization Across Developmental Stages Next, we assessed generalization across the lifespan by testing on HCP-D (children and adolescents) and HCP-A (aging adults). On HCP-D, BOLD-Cast exhibited the largest error increase ($MAE = 21.68$), yet still retained competitive structural consistency ($FC_{PCC} = 0.98$). On HCP-A, BOLD-Cast maintained low error ($MAE = 17.79$, $MAPE = 0.23$) and high correlation consistency ($mPCC = 0.99$), indicating that despite age-related neural structural changes in the aging brain, the macroscopic dynamic features remain relatively stable. These results reflect the model's adaptability to gradual (aging) and nonlinear (developmental) neural changes, as well as its ability to track functional organization in non-adult cohorts.

Generalization Across Clinical Conditions Finally, we evaluated BOLD-Cast on ABIDE to test the model's ability to generalize to clinical and behavioural variability. ABIDE comprises a heterogeneous group of subjects with ASD and typically developing controls (TC), exhibiting significant heterogeneity in FC patterns. While performance slightly declined compared to HCP-YA ($MAE = 7.31$, $MAPE = 0.61$), BOLD-Cast still surpassed all baselines in all metrics and maintained strong structural fidelity ($FC_{PCC} = 0.98$). Notably, although inter-subject variability

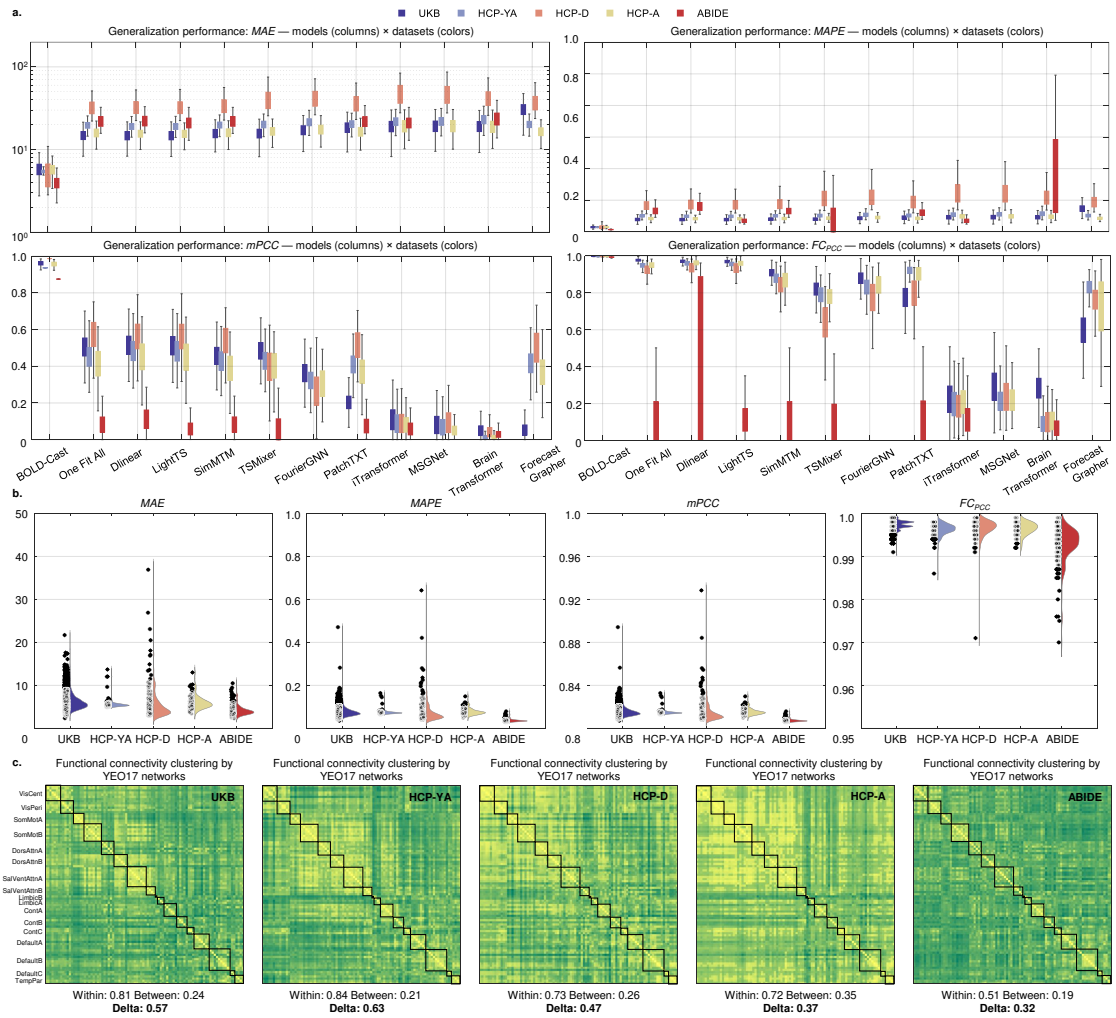


Fig. 4 Cross-dataset generalization performance of BOLD-Cast. **a**, Boxplots of subject-level prediction metrics (MAE, $MAPE$, $mPCC$, and FC_{PCC}) for BOLD-Cast and 11 baseline models across five datasets: UKB, HCP-YA, HCP-A, HCP-D, and ABIDE. Box colors indicate dataset origin; models are grouped by columns. Due to a high incidence of NaN values during validation on ABIDE, some models are not represented in the corresponding boxplots. BOLD-Cast consistently achieves lower prediction errors and higher functional preservation across datasets, particularly in UKB and HCP-YA. Error variance increases in HCP-D, potentially due to higher inter-subject variability in the adolescent population. **b**, Raincloud plots visualizing subject-level prediction distributions for BOLD-Cast. MAE and $MAPE$ remain low across datasets, while $mPCC$ and FC_{PCC} stay consistently high, indicating strong individual-level predictive fidelity. **c**, Group-averaged FC matrices generated from BOLD-Cast predictions using the CC200 parcellation, visualized with Yeo-17 network boundaries. Within- and between-network similarity scores are reported (e.g., Within: 0.51, Between: 0.19 in ABIDE; $\Delta = 0.32$). All datasets preserve canonical resting-state modularity, confirming that shared latent representations learned by BOLD-Cast can generalize across populations while maintaining neurobiologically meaningful architecture.

was larger in ABIDE (Fig. 4b), the model preserved functional topology across clinical subgroups. This result demonstrates its potential for robust transfer across unknown pathological conditions.

Together, these findings confirm that BOLD-Cast generalizes effectively across acquisition protocols, developmental stages, and clinical conditions without adaptation. This zero-shot robustness, backed by both temporal accuracy and topological consistency, underscores its potential for broad deployment in heterogeneous clinical imaging scenarios.

2.4 Interpreting the Mechanism Underlying BOLD-Cast's Generalization

Building on the previous results demonstrating BOLD-Cast's generalization across diverse cohorts, we further explored the mechanism basis underlying its generalization ability. Specifically, we aimed to determine whether the cohort-invariant embeddings $\mathbf{F}_c$ extracted by the

model reflected intrinsic brain connectomes that are conserved across populations. To this end, we reorganized FC derived from $\mathbf{F}_c$ based on canonical resting-state networks using the established mapping between CC200 and the Yeo-17 atlas [55]. This ensured that our interpretability analysis remained grounded in the CC200 framework and stayed consistent with the prediction task.

As shown in Fig. 4**c**, the connectivity patterns extracted by BOLD-Cast consistently exhibit well-defined resting-state networks (such as default mode, dorsal attention, salience, and control networks), with within-network connectivity systematically stronger than between-network connectivity. This spatial organization is widely considered a hallmark of intrinsic functional brain activity, confirming that the common representations encode biologically meaningful, domain-invariant connectomes rather than spurious statistical correlations. These findings provide a mechanistic explanation for BOLD-Cast's generalization capacity and support its robustness for real-world clinical deployment. Visualization results of several subjects are shown in the supplementary material.

2.5 BOLD-Cast-Predicted fMRI Retains Discriminative Signal for Clinical-Relevant and Cognitive Modeling

Building on the BOLD-Cast's high prediction accuracy and robust generalization, we next evaluated whether its generated fMRI data retain discriminative utility in clinical and cognitive downstream applications. This analysis addresses a core question: Does generated fMRI time-series preserve the discriminative signal needed for downstream neurocognitive and clinical-related modeling?

We selected three representative downstream tasks: ASD classification [56–58], sex identification, and cognitive score prediction [59–61], reflecting typical modeling scenarios involving clinical heterogeneity, group-level distinction, and individual variation respectively. Evaluations were conducted on ABIDE and HCP-YA using BrainGB [62] and PTGB [63], two recent graph neural networks that have been widely adopted for fMRI downstream task. Both BOLD-Cast and One Fit All served as upstream generators, where short real fMRI time series were used for training and predicted sequences were used for testing.

For the ABIDE dataset, we first used the initial 150 TRs (5 minutes, TR=2000ms) of real fMRI data as input to the upstream generators, which predicted the subsequent 150 TRs. The downstream models (BrainGB/PTGB) were trained exclusively on full-length real fMRI time series from the training set. For testing, we used extended sequences composed of 150 real TRs (same as the input segment) plus 150 predicted TRs, forming a 450-TR virtual sequence. For the HCP-YA dataset, the experimental setup followed the same logic but with TR-specific parameters adjusted to the dataset: we used the initial 415 TRs ($\approx$5 minutes) of real fMRI data as input to generate the next 415 TRs. Downstream models for sex identification and cognition predicition were again trained on real full-length fMRI data from the training set, and tested on extended sequences (415 real TRs + 415 predicted TRs).

Results (Fig. 5**a**) show that BOLD-Cast consistently outperforms One Fit All in all tasks. In the ABIDE cohort, BOLD-Cast achieved superior accuracy, precision, and $F1$ score, indicating a stronger capacity to recover biomarker-linked FC differences known to relate to ASD risk stratification. In HCP-YA, BOLD-Cast also achieved advantages in MAE, $RMSE$, and PCC metrics, particularly in the Flanker and PicSeq tasks, suggesting a better preservation of individual cognitive fingerprints. To further elucidate these findings, we examined the FC maps and cortical error distributions derived from real and predicted fMRI time series (Fig. 5**b**), with visualizations generated using ParaView. For FC map construction, we first represented the cerebral cortex as a triangular mesh in ParaView—each triangle defined by connecting three cortical vertices, and each brain region by a central point. FC edges were then modelled as lines linking these regional central points. We quantified discrepancies between FC matrices from real and BOLD-Cast-predicted fMRI time series: edges with a deviation > 0.3 were marked in red, while those with a deviation ≤ 0.3 were marked in black (left panel of Fig. 5**b**). For cortical error distribution visualization (right panel of Fig. 5**b**), we mapped the voxel-wise deviation between predicted and real fMRI signals onto the triangular cortical mesh. Here, red indicated larger deviations and blue indicated smaller deviations. BOLD-Cast-predicted FC errors (red edges) were significantly fewer and more spatially localized. Importantly, in ABIDE, both real and BOLD-Cast-predicted FCs exhibited the well-known fronto-posterior hypoconnectivity pattern—an established biomarker

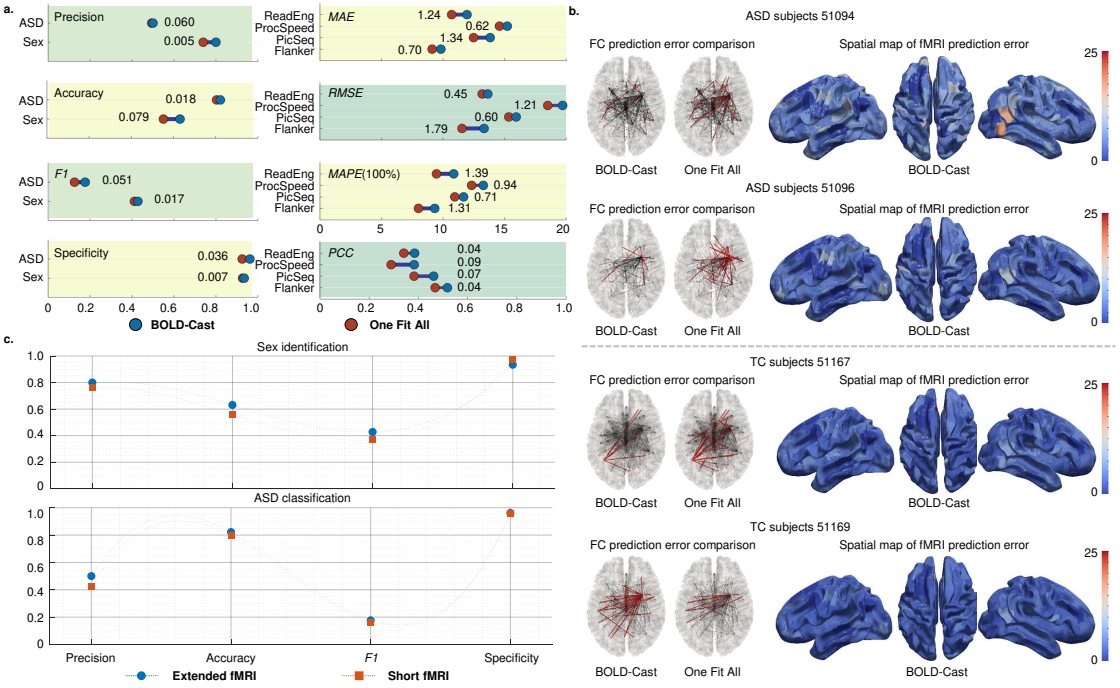


Fig. 5 Downstream task performance enabled by BOLD-Cast-predicted fMRI. **a**, Comparison of downstream performance using fMRI predicted by BOLD-Cast (blue) and One Fit All (red) across two types of tasks: classification (ASD, Sex) and four cognitive prediction tasks (Reading Comprehension, Processing Speed, Picture Sequence, Flanker). The numbers adjacent to the lollipop indicate the performance difference for each metric. BOLD-Cast achieving superior precision, accuracy, F1, and specificity across all classification tasks, alongside lower MAE, $RMSE$, and $MAPE$ in regression. Notably, BOLD-Cast-predicted fMRI achieves higher discriminative power even under identical input durations. **b**, Error decomposition of BOLD-Cast predictions. The left panels visualize FC comparisons (with correctly predicted FCs marked in black and incorrectly predicted ones in red). The right panels map the spatial distribution of the absolute mean prediction error for each brain parcel's predicted fMRI time series (blue = smaller prediction deviation, red = larger deviation). This decomposition shows that BOLD-Cast's superior downstream performance stems from both higher FC accuracy and more anatomically localized temporal prediction errors. **c**, Slop chart of downstream classification performance. BOLD-Cast consistently outperforms One Fit All in ASD and sex tasks across all evaluated metrics, with larger margins in precision and accuracy. These results underscore the utility of high-fidelity generated fMRI in supporting clinically relevant predictions, particularly in scenarios where only limited scan durations are available.

of ASD [64, 65]. This indicates that the predicted sequences preserve biologically meaningful connectivity organization, not merely short-term temporal smoothness.

2.6 Temporal Extension Improves FC Stability and Predictive Robustness

In real-world clinical scenarios, particularly involving children or patients with ASD, fMRI acquisition is often time-constrained due to motion or tolerance limits, which weakens the stability of FC estimates and degrades downstream predictive modeling. To examine whether BOLD-Cast can mitigate this limitation under a fixed acquisition time, we conducted a controlled experiment on ABIDE: for each subject, we used the latter 150 TRs (TRs 151–300, corresponding to 5 minutes, TR=2000ms) from the full set of 300 real fMRI TRs (10 minutes) as input to BOLD-Cast. The model then generated 150 consecutive virtual timepoints that extend beyond the original real sequence. We appended these 150 predicted TRs to the initial 300 real TRs, forming an extended 450-TR sequence (15 minutes). We then evaluated whether this virtual "longer-duration" sequence improves robustness in two representative downstream predictive tasks: ASD classification and sex identification, both implemented using BrainGB.

As shown in Fig. 5**c**, fMRI time series extended by BOLD-Cast (blue lines) consistently outperformed or closely matched the original sequences (red lines) across all four metrics: In ASD classification, the extended sequences achieved higher accuracy (+0.02) and $F1$ score (+0.03), while maintaining strong precision and specificity. It is worth noting that in binary classification, there is often a trade-off between sensitivity (recall) and specificity. The observed improvement

in F1 score—which incorporates recall—without a reduction in specificity indicates that the temporally extended sequences not only enhance the detection of true positives but also preserve the correct identification of true negatives, a critical balance for clinical utility. This indicates that the virtual longer duration preserved group-level discriminative signal. In sex identification, minor decreases (≤ 0.01) in some metrics were observed, yet overall accuracy slightly improved, confirming that sequence extension did not introduce systematic bias or degrade group-level separability. The improvement is attributable not to simply "adding more points", but to stabilizing functional connectivity patterns over a longer temporal horizon, thereby providing a more reliable basis for downstream predictive modeling under constrained acquisition budgets.

Overall, these findings demonstrate that model-based temporal extension can act as a surrogate for longer acquisitions, improving the stability of functional features available to downstream predictive models without increasing scan duration.

3 Discussion

We introduce BOLD-Cast, an autoregressive fMRI data generation framework, to address the long-standing challenge of scan time constrains in clinical neuroimaging. BOLD-Cast leverages temporal modeling and a dual-stream representation strategy that disentangles cohort-invariant and subject-specific embeddings. Compared to existing models that primarily focus on short-term, local predictions or fixed data regimes, BOLD-Cast emphasizes a balance between modeling intrinsic neural dynamics and ensuring adaptability across heterogeneous populations. This design allows the model to maintain temporal coherence while effectively capturing biological connectivity patterns, achieving high-fidelity construction of long-range brain dynamics from short observations. Therefore, beyond demonstrating accurate spatiotemporal consistency in prediction task, BOLD-Cast achieves remarkable performance in zero-shot testing across multiple datasets and populations. This design bridges a critical gap in current fMRI prediction methods, which often trade-off between generalization capacity, physiological plausibility, and downstream task utility. Moreover, the fMRI time series generated by BOLD-Cast were shown to support a range of downstream tasks, including ASD classification, sex identification, and cognitive score prediction, while also enhancing both classification performance without additional acquisition costs. Collectively, these findings position BOLD-Cast as a flexible and generalizable solution for efficient brain function assessment in time-constrained clinical settings.

BOLD-Cast's robust generalization and high flexibility across populations and tasks suggests strong potential for deployment in clinical scenarios characterized by limited patient cooperation. In such settings, scan durations are often constrained by attentional limitations, anxiety, or motion-related artifacts, particularly common among children, the elderly, or cohorts with neuropsychiatric conditions, leading to unstable connectivity estimates and reduced diagnostic reliability. BOLD-Cast can synthesize longer brain dynamic sequences from existing short time series to improve the stability of FC and the discriminative power of downstream analyses. For example, in the ABIDE dataset comprising pediatric ASD patients, the expanded fMRI time series generated by BOLD-Cast enhanced diagnostic accuracy and individual discrimination without introducing systematic biases (i.e., without shifting group-level class balance or globally inflating/deflating FC magnitude). Given that scans in this population are often interrupted, BOLD-Cast supports the feasibility of a "minimal acquisition + model completion" strategy that reduces dependence on prolonged scanning. Furthermore, its independence from task paradigms or behavioural responses makes it more adaptable to non-cooperative populations. As a general-purpose data extension module, BOLD-Cast could be integrated into existing acquisition protocols to improve connectome stability, lower scan failure rates, and facilitate the broader clinical translation of fMRI for early-stage assessments in neurodevelopmental and psychiatric conditions.

While BOLD-Cast demonstrates robustness across multiple datasets and downstream tasks, the present study primarily focused on resting-state fMRI for validation. Compared to resting-state data, task-based fMRI (task-fMRI) can provide stimulus-evoked, causally interpretable neural responses that are particularly valuable for modeling cognitive processes and disease mechanisms. However, task-fMRI data acquisition often relies on complex task design and high subject compliance, which currently limit its feasibility in routine clinical practice. For this reason, our evaluation prioritized resting-state fMRI as the most accessible and widely adopted

modality, allowing for a systematic assessment of model robustness under clinically relevant, task-independent conditions. It is worth noting that BOLD-Cast's modeling framework is inherently scalable and paradigm-agnostic, suggesting that future studies could investigate its applicability to task-based settings to further enhance its interpretability and potential for functional decoding in specific cognitive or interventional contexts.

Future studies may extend the BOLD-Cast framework along several clinically relevant applications. First, the model could be evaluated on task-based fMRI data to assess its ability to capture stimulus-evoked dynamics, predict task-specific cognitive functions, and support post-intervention monitoring. Such extension would be particularly valuable in pediatric neurodevelopment, attention disorders, and other domains where functional evaluation relies on task engagement. Second, integration with other neurophysiological modalities such as electroencephalography (EEG) or functional near-infrared spectroscopy (fNIRS) may enhance the model's sensitivity to latent neural states in short acquisition windows and improve cross-modal consistency. These multimodal extensions could not only reinforce generalizability and physiological interpretability, but also lay the groundwork for non-invasive neuromodulation and hybrid brain–computer interface applications. Finally, to support real-world clinical deployment, further work is needed to examine model robustness across a broader range of acquisition protocols, scanner models, and hardware configurations. Building low-latency, interpretable, and end-to-end generative–discriminative systems will be critical to meeting the reliability and safety standards required in practical clinical workflows.

4 Methods

4.1 Overview of the Framework

BOLD-Cast, proposed in this study, is a two-stage framework for brain fMRI time series prediction (Fig. 6). It consists of two modules: pretrained graph disentanglement module and LLM-driven time series modeling module. These two modules are designed to extract interpretable network representations and model the evolution of brain dynamics across time.

In Stage I, we pre-train a graph autoencoder on the input fMRI sequence $x_{1:h}$, aiming to learn stable and biologically meaningful connectivity connectomes. Each sliding-window sequence is encoded into two complementary representations: a cohort-invariant embedding $\mathbf{F}_C$ capturing shared functional patterns across subjects, and a unique subject-specific embedding $\mathbf{F}_P$ reflecting private variations. In Stage II, we introduce temporal prompts to enhance the model's sensitivity to time-dependent dynamics. Specifically, the extracted $\mathbf{F}_C$ and $\mathbf{F}_P$ embeddings, along with the history sequence S and timestamp-based text embedding $\mathbf{F}_T$, are concatenated and processed by a frozen GPT-2 model to autoregressively predict future brain activations $\hat{x}_{h+1:h+p}$. At each step, the model generates the next neural activation vector conditioned on the historical context.

This architecture effectively integrates functional connectivity connectomes with temporal prompts, enabling BOLD-Cast to preserve inter-individual diversity while enhancing the modeling of shared neural dynamics over time.

4.2 Stage I: Decoupled Representation of Common and Private Information

This stage aims to extract disentangled representations from fMRI data by modeling both shared (common) and individual (private) brain connectivity patterns. It consists of three parts: (i) constructing subject-specific graphs from fMRI time series, (ii) extracting common representations across graphs, and (iii) refining private representations for individual-specific variation modeling.

4.2.1 Graph Construction and Node Embedding

Given fMRI time series data, we first segment each subject's recording into overlapping windows of 81 timepoints (corresponding to $\approx$ 60 seconds in UKB, $TR = 735ms$). The sliding window advances with step size = 1 TR. For each window, we compute the FC matrix by calculating the Pearson correlation between region time series. The resulting FC matrices are binarized using a top-80% K-nearest neighbor (KNN) strategy [66] to construct sparse and personalized adjacency matrices $\mathbf{A}^r$ for each subject $r = 1, \ldots, R$. Each graph thus captures the subject-specific functional

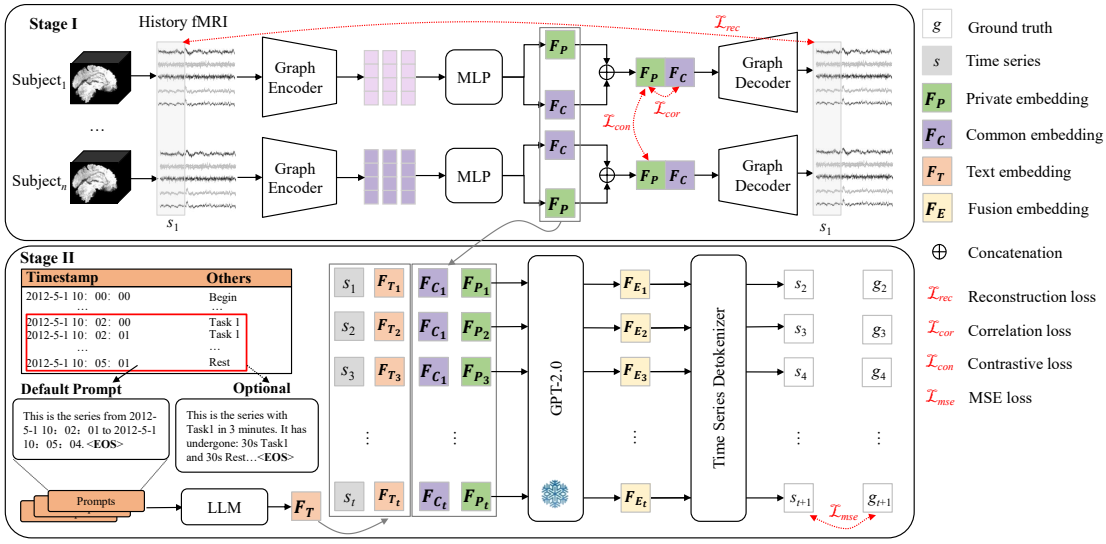


Fig. 6 Overview of BOLD-Cast. Stage I: disentangled graph representation learning separates common and individual FC patterns, enforcing alignment of cohort-invariant embeddings and suppression of noise in individual embeddings. Stage II: timestamp-based textual prompts are embedded by a frozen GPT-2 model, concatenated with graph embeddings, and processed by an LLM to predict temporally extended fMRI signals.

topology within a given temporal window. Preprocessing procedures of fMRI to get parcel time series data for each dataset are provided in Supplementary Table S8.

Each subject's graph $\mathcal{G}^r = (\mathbf{A}^r, \mathbf{S}^r)$ is defined by the adjacency matrix and node features $\mathbf{S}^r$ (e.g., parcel-level activation in the same window). Node embeddings $\mathbf{H}^r$ are obtained via graph convolutional networks (GCNs) [67, 68]:

$$\mathbf{H}^r = \sigma(\hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}^r\hat{\mathbf{D}}^{-1/2}\mathbf{S}^r\Theta), \tag{1}$$

where $\hat{\mathbf{A}}^r = \mathbf{A}^r + \omega\mathbf{I}$ includes self-loops, $\hat{\mathbf{D}}$ is the degree matrix, Θ is the GCN weight matrix, and $\sigma(\cdot)$ is a non-linear activation function.

4.2.2 Decoupling Common and Private Information

To disentangle the shared and idiosyncratic components of brain connectivity, the learned node embeddings $\mathbf{H}^r$ are passed through two parallel multilayer perceptron (MLP) [69] modules to extract subject-wise common $\mathbf{F}_C^r$ and private $\mathbf{F}_P^r$ representations. The goal is to achieve (1) aligned, complete common representations across subjects and (2) uncorrelated clean separation between $\mathbf{F}_C^r$ and $\mathbf{F}_P^r$.

Common Information Alignment. To avoid simply averaging all $\mathbf{F}_C^r$ and losing discriminative power, we introduce a shared latent variable $\mathbf{L}_S$ with orthogonality and zero-mean constraints via singular value decomposition (SVD). A matching loss aligns each subject's common representation $\mathbf{F}_C^r$ with the shared variable $\mathbf{L}_S$:

$$L_{\text{match}} = \frac{1}{N}\sum_{r=1}^{R}\sum_{n=1}^{N}(c_n^r - ls_n)^2, \quad \text{s.t. } \frac{1}{N}\sum_n ls_n ls_n^\top = \mathbf{I}, \quad \frac{1}{N}\sum_n ls_n = \mathbf{0}. \tag{2}$$

Common-Private Separation. To enhance decoupling, we minimize the linear correlation between the common and private embeddings by computing the Pearson correlation between the outputs of the respective MLPs:

$$L_{\text{pear}} = \sum_{r=1}^{R}\frac{|\text{Cov}(\phi(\mathbf{F}_C^r), \varphi(\mathbf{F}_C^r))|}{\sqrt{\text{Var}(\phi(\mathbf{F}_C^r))}\sqrt{\text{Var}(\varphi(\mathbf{F}_C^r))}}, \tag{3}$$

where $\phi(\cdot)$ and $\varphi(\cdot)$ denote the learned transformations from the MLPs.

4.2.3 Refinement of Private Information

While common representations capture shared connectivity, private representations $\mathbf{F}_P^r$ are intended to model subject-specific or window-specific variations. However, these may also contain noise. Formally, let $\mathcal{E}^r$ denote the set of all edges in graph $\mathcal{G}^r$. We first focus on the individual-specific edges of $\mathcal{G}^r$, i.e., edges $e_{ij}^r \in \mathcal{E}^r$ that do not appear in any other graph's edge set (i.e., $e_{ij}^r \notin \bigcup_{r' \in [1,R], r' \neq r} \mathcal{E}^{r'}$, among R total graphs). Based on the cosine similarity of their node pairs in the shared latent variable $\mathbf{L}_S$, we define two subsets of these individual-specific edges for $\mathcal{G}^r$:

- $\mathcal{E}_c^r$: Set of complementary edges in $\mathcal{G}^r$, consisting of individual-specific edges whose connected node pairs have high cosine similarity in $\mathbf{L}_S$;
- $\mathcal{E}_n^r$: Set of noisy edges in $\mathcal{G}^r$, consisting of individual-specific edges whose connected node pairs have low cosine similarity in $\mathbf{L}_S$.

We then apply a contrastive learning objective to push complementary edges closer in the private space, and noisy edges further apart:

$$L_{\text{con}} = -\sum_{r=1}^{R} \log \frac{\sum\limits_{e_{i,j} \in \mathcal{E}_c^r} \exp\left(\theta(p_i^r, p_j^r)/\tau\right)}{\sum\limits_{e_{i,j} \in \mathcal{E}_c^r} \exp\left(\theta(p_i^r, p_j^r)/\tau\right) + \sum\limits_{e_{k,l} \in \mathcal{E}_n^r} \exp\left(\theta(p_k^r, p_l^r)/\tau\right)}, \tag{4}$$

where p_i, p_j, p_k, and p_l denote the individual representations of nodes v_i, v_j, v_k, and v_l, respectively, with $(v_i, v_j) = e_{ij}^r \in \mathcal{E}_c^r$ and $(v_k, v_l) = e_{kl}^r \in \mathcal{E}_n^r$. Additionally, θ represents the cosine similarity operation, and τ is a temperature parameter. The positive and negative edges are dynamically sampled based on top-k similarity heuristics derived from $\mathbf{L}_S$.

Finally, a reconstruction loss is applied to ensure the representations retain the original node features and local structure. The combined embeddings $[\mathbf{F}_C^r; \mathbf{F}_P^r]$ are used to reconstruct both:

$$L_{\text{recon}} = \frac{1}{N} \sum_{r=1}^{R} \sum_{n=1}^{N} \left(\|\bar{x}_n - x_n\|^2 + \|\bar{x}_n - x_{n,\text{nei}}\|^2 \right), \tag{5}$$

where $x_{n,\text{nei}} = \frac{1}{m} \sum_{j=1}^{m} x_j,\ v_j \in \mathcal{N}_n$ is the average of m sampled neighbors.

4.3 Stage II: Prompt-Based Temporal Modeling with Language Embedding

To capture the temporal dynamics of brain activity beyond static graph features, we propose a prompt-driven autoregressive modeling framework via GPT-2. This stage is designed to predict futurefMRI using FC embeddings derived from Stage I outputs and contextual information encoded as natural language. When selecting GPT-2 as the core model for our time-series prediction task, we primarily considered its open-source customizability and data privacy protection capabilities. Unlike closed-source large models (such as ChatGPT-5 [70], Grok [71], Gemini [72], etc.), GPT-2's model weights, code, and tokenizer are publicly available, allowing us to modify the network architecture or add adaptation layers to meet the specific needs of the medical field and research requirements. This flexibility is crucial for ensuring both the adaptability and scalability of the model. Furthermore, GPT-2 supports local inference, eliminating the need to upload sensitive data such as timestamps or brain function data to cloud-based APIs, which aligns with compliance requirements in privacy-sensitive domains such as healthcare and finance, thereby providing an additional layer of security for our application. Compared to larger or more complex LLMs, GPT-2 is computationally efficient, offering fast inference and fine-tuning even under limited computational resources, making it particularly suitable for resource-constrained practical scenarios [73, 74]. Additionally, GPT-2's tokenizer and positional encoding seamlessly handle multi-modal data such as text, timestamps, and functional connectivity matrices, enabling cross-modality alignment and significantly enhancing the representation of temporal features [75, 76]. Considering these advantages, GPT-2 presents an ideal and feasible solution for our time-series prediction tasks in the field of brain science.

Tokenization of Temporal Inputs: For each sliding window of h time steps (e.g., 81 TR corresponding to 1 minute in fMRI), we construct a sequence of input tokens for the LLM. At each

time step t, a token is composed of four components concatenated along the feature dimension: the shared representation $\mathbf{F}_C^r \in \mathbb{R}^{1\times 128}$, the subject-specific representation $\mathbf{F}_P^r \in \mathbb{R}^{1\times 128}$, the raw fMRI embedding $\mathbf{F}_S^r$, and the temporal prompt embedding $\mathbf{F}_T^r$. The full token at each time step is thus $z_t = \mathbf{F}_C^r \oplus \mathbf{F}_P^r \oplus S \oplus \mathbf{F}_T^r$, and the input sequence is $\mathbf{Z} = [z_1, \ldots, z_h]$.

Temporal Prompt Embedding: Instead of relying on standard positional encodings, BOLD-Cast introduces natural language prompts to encode temporal context. For rsfMRI, each sliding window corresponds to a concrete time range (e.g., "2017-07-01 09:00:00 to 09:01:00"). These timestamp descriptions are converted into text prompts and encoded via a frozen GPT-2 model, yielding the temporal embedding $\mathbf{F}_T^r$ for each time step:

$$\mathbf{F}_T^t = \text{SelectLast}(\text{GPT-2}(\text{PromptText}(t))). \tag{6}$$

Autoregressive Forecasting with Frozen LLM: The input sequence $\mathbf{Z}$ is passed through a pre-trained and frozen GPT-2 model to model sequential dependencies. A linear adapter layer is inserted before the input layer to project z_t into the token embedding space. The model generates future fMRI vectors $\hat{x}_{h+1:h+p}$ autoregressively for p steps, using a projection head to map GPT-2 outputs back to the original fMRI space.

4.4 Training Objectives and Optimization Strategy

Stage I Training. The graph encoder is pre-trained in a self-supervised manner using a combination of objectives to disentangle common and private representations of brain connectivity:

$$L_{\text{Stage-I}} = \lambda_1 L_{\text{match}} + \lambda_2 L_{\text{pear}} + \lambda_3 L_{\text{recon}} + \lambda_4 L_{\text{con}}. \tag{7}$$

The encoder is trained until convergence and frozen thereafter. Detailed formulations are given in the previous section.

Stage II Training. Once Stage I is complete, the extracted common ($\mathbf{F}_C^{1:h}$) and private ($\mathbf{F}_P^{1:h}$) representations are cached and used as structural priors. These are combined with raw signal embeddings $S^{1:h}$ and timestamp-based text embeddings $\mathbf{F}_T^{1:h}$ to form the autoregressive input sequence $Z = [z_1, \ldots, z_h]$ for the LLM.

During training, we adopt a non-teacher-forcing autoregressive strategy. For each training instance, the model observes a historical window of h time steps (e.g., $h = 81$) and predicts the subsequent p steps (e.g., $p = 81$), yielding outputs $\hat{x}_{h+1:h+p} \in \mathbb{R}^{p\times N}$ where N is the number of regions. A projection head $\psi(\cdot)$ is applied to map LLM output tokens back to the fMRI signal space:

$$\hat{x}_{h+1:h+p} = \psi(\text{LLM}(Z)), \tag{8}$$

where ψ consists of a linear layer and layer norm.

The training loss is defined as *MSE* between predicted and real fMRI signals:

$$L_{\text{Stage-II}} = \frac{1}{pN} \sum_{t=h+1}^{h+p} \sum_{n=1}^{N} \|\hat{x}_{t,n} - x_{t,n}\|^2. \tag{9}$$

Model parameters are updated using the Adam optimizer with early stopping based on validation *mPCC*. No teacher forcing is applied during training or inference, enabling long-horizon, autoregressive forecasting based solely on past context and temporal prompts.

Ablation Study. To examine the contributions of each module, we conducted ablation experiments including removal of common/private disentanglement, replacement of text prompts with positional encodings, and masking of timestamp embeddings. Results are reported in Supplementary Figure 5.

4.5 Implementation Details

Environment and Training. All experiments were implemented in PyTorch and run on NVIDIA V100 GPUs (32 GB memory). Model training used a batch size of 100 and the Adam optimizer with a learning rate of 3×10^{-4}. Training was performed for up to 100 epochs with early stopping (patience = 10) based on validation *mPCC*. The historical context window was set to $h = 81$ and the prediction horizon to $p = 81$.

Two-Stage Training. BOLD-Cast adopts a two-stage training pipeline. In Stage I, the graph encoder is trained independently to extract disentangled representations of functional connectivity, including cohort-invariant embedding $\mathbf{F}_C$ and private embedding $\mathbf{F}_P$. Once trained, this module is frozen. In Stage II, the downstream GPT-2 forecaster is trained with *MSE* loss to model long-horizon brain dynamics, using only the fixed embeddings from Stage I and timestamp-aware prompt tokens as input. No teacher forcing is applied during prediction.

Model Architecture. The graph encoder consists of a two-layer GCN followed by a 2-layer MLP for both $\mathbf{F}_C$ and $\mathbf{F}_P$ branches, with tanh activation and 0.1 dropout. Each embedding is projected to a 128-dimensional latent space. For forecasting, the GPT-2 backbone is frozen and augmented with a lightweight adapter and projection head to decodefMRI time. Detailed configurations, including layer dimensions, adapter design, and learning rate schedules for all baselines and BOLD-Cast variants, are provided in Supplementary Tables 9–14.

Downstream Evaluation. Phenotype regression and classification tasks are conducted using predictedfMRI time series. We extract functional connectivity matrices from the model outputs and feed them into BrainGB or PTGB classifiers for downstream performance assessment. All post-hoc evaluation protocols and metrics are detailed in Supplementary Tables 15–16.

Data Availability

The data analysed in this study are available for bona fide research purposes. We have obtained the appropriate data access permissions for each data source. Due to privacy and ethical considerations, the data cannot be deposited in a public repository. Researchers interested in accessing the data should directly contact the responsible authorities managing each data source to obtain access permissions:

- UK Biobank: (https://www.ukbiobank.ac.uk/)
- Autism Brain Imaging Data Exchange (http://fcon_1000.projects.nitrc.org/indi/abide/)
- Human Connectome Project–Young Adult: (https://www.humanconnectome.org/study/hcp-young-adult)
- Human Connectome Project–Development: (https://www.humanconnectome.org/study/hcp-lifespan-development)
- Human Connectome Project–Aging: (https://www.humanconnectome.org/study/hcp-lifespan-aging)

No new data were collected as part of this study. All analyses were conducted on pre-existing publicly accessible datasets with appropriate approvals.

Code Availability

The PyTorch code for the BOLD-Cast proposed in this study is available at the GitHub repository here, including installation instructions, model weights, and evaluation code.

Contributors

Y.J. and Y.Y. conceived the study and proposed the overall methodology. Y.J. implemented the core model, conducted experiments, performed result interpretation, and drafted the manuscript. C.W. contributed to the large-scale UK Biobank data processing and provided key preprocessing pipelines. C.W. and Z.H. performed data preprocessing across all datasets. Y.S. participated in experimental analysis and interpretation of downstream results. Z.P. contributed to manuscript writing and revisions. T.Z. and P.Y.M.W. provided critical feedback on analysis and experimental results. J.L. and Y.Y. reviewed the manuscript and supervised the research process. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Organization, W.H.: World Mental Health Today: Latest Data. World Health Organization, Geneva (2025)

[2] Jiao, Y., Zhao, K., Wei, X., Carlisle, N.B., Keller, C.J., Oathes, D.J., Fonzo, G.A., Zhang, Y.: Deep graph learning of multimodal brain networks defines treatment-predictive signatures in major depression. Molecular Psychiatry, 1–12 (2025)

[3] Zhao, K., Xie, H., Fonzo, G.A., Tong, X., Carlisle, N., Chidharom, M., Etkin, A., Zhang, Y.: Individualized fmri connectivity defines signatures of antidepressant and placebo responses in major depression. Molecular psychiatry **28**(6), 2490–2499 (2023)

[4] Berto, S., Treacher, A.H., Caglayan, E., Luo, D., Haney, J.R., Gandal, M.J., Geschwind, D.H., Montillo, A.A., Konopka, G.: Association between resting-state functional brain connectivity and gene expression is altered in autism spectrum disorder. Nature Communications **13**(1), 3328 (2022)

[5] Norman, L.J., Sudre, G., Bouyssi-Kobar, M., Jiao, M., Gligorovic, S., Jean, J., White, T., Shaw, P.: Cross-sectional mega-analysis of resting-state alterations associated with autism and attention-deficit/hyperactivity disorder in children and adolescents. Nature Mental Health, 1–15 (2025)

[6] Ward, N., Brown, M., Thompson, A., Frackowiak, R.: Neural correlates of motor recovery after stroke: a longitudinal fmri study. Brain **126**(11), 2476–2496 (2003)

[7] Murphy, T.H., Corbett, D.: Plasticity during stroke recovery: from synapse to behaviour. Nature reviews neuroscience **10**(12), 861–872 (2009)

[8] Duncan, J.S.: Imaging in the surgical treatment of epilepsy. Nature Reviews Neurology **6**(10), 537–550 (2010)

[9] Lucas, A., Revell, A., Davis, K.A.: Artificial intelligence in epilepsy—applications and pathways to the clinic. Nature Reviews Neurology **20**(6), 319–336 (2024)

[10] Mhanna, H.Y.A., Omar, A.F., Radzi, Y.M., Oglat, A.A., Akhdar, H.F., Al Ewaidat, H., Almahmoud, A., Yaseen, A.-B.B., Al Badarneh, L., Alhamad, O., et al.: Systematic review of functional magnetic resonance imaging (fmri) applications in the preoperative planning and treatment assessment of brain tumors. Heliyon **11**(3) (2025)

[11] Cash, R.F., Cocchi, L., Lv, J., Wu, Y., Fitzgerald, P.B., Zalesky, A.: Personalized connectivity-guided dlpfc-tms for depression: Advancing computational feasibility, precision and reproducibility. Human brain mapping **42**(13), 4155–4172 (2021)

[12] Cash, R.F., Cocchi, L., Lv, J., Fitzgerald, P.B., Zalesky, A.: Functional magnetic resonance imaging–guided personalization of transcranial magnetic stimulation treatment for depression. JAMA psychiatry **78**(3), 337–339 (2021)

[13] Birn, R.M., Molloy, E.K., Patriat, R., Parker, T., Meier, T.B., Kirk, G.R., Nair, V.A., Meyerand, M.E., Prabhakaran, V.: The effect of scan length on the reliability of resting-state fmri connectivity estimates. Neuroimage **83**, 550–558 (2013)

[14] Kumar, V., Lee, J., Liu, H.-L., Allen, J., Filippi, C., Holodny, A., Hsu, K., Jain, R., McAndrews, M., Peck, K., *et al.*: Recommended resting-state fmri acquisition and preprocessing steps for preoperative mapping of language and motor and visual areas in adult and pediatric patients with brain tumors and epilepsy. American Journal of Neuroradiology **45**(2), 139–148 (2024)

[15] Ooi, L.Q.R., Orban, C., Zhang, S., Nichols, T.E., Tan, T.W.K., Kong, R., Marek, S., Dosenbach, N.U., Laumann, T.O., Gordon, E.M., et al.: Longer scans boost prediction and cut

costs in brain-wide association studies. Nature, 1–10 (2025)

[16] Tanrıtanır, A.C., Villringer, K., Galinovic, I., Grittner, U., Kirilina, E., Fiebach, J.B., Villringer, A., Khalil, A.A.: The effect of scan length on the assessment of bold delay in ischemic stroke. Frontiers in Neurology **11**, 381 (2020)

[17] Lawal, O., Regelous, P., Omiyi, D.: Supporting claustrophobic patients during magnetic resonance imaging examination–the patient perspective. Radiography **29**(6), 1108–1114 (2023)

[18] Van Dijk, K.R., Sabuncu, M.R., Buckner, R.L.: The influence of head motion on intrinsic functional connectivity mri. Neuroimage **59**(1), 431–438 (2012)

[19] Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E.: Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. Neuroimage **59**(3), 2142–2154 (2012)

[20] Wilson, S., Fan, C.C., Hewitt, J.: Abcd behavior genetics: Twin, family, and genomic studies using the adolescent brain cognitive development (abcd) study dataset. Behavior genetics **53**(3), 155–158 (2023)

[21] Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., *et al.*: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry **19**(6), 659–667 (2014)

[22] Griswold, M.A., Jakob, P.M., Heidemann, R.M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., Haase, A.: Generalized autocalibrating partially parallel acquisitions (grappa). Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **47**(6), 1202–1210 (2002)

[23] Pruessmann, K.P., Weiger, M., Scheidegger, M.B., Boesiger, P.: Sense: sensitivity encoding for fast mri. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **42**(5), 952–962 (1999)

[24] Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., Uğurbil, K.: Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. Magnetic resonance in medicine **63**(5), 1144–1153 (2010)

[25] Wall, M.B.: Multiband acquisition sequences for fmri: Proceed with caution. Aperture Neuro **3** (2023)

[26] Ghobadi-Azbari, P., Jamil, A., Yavari, F., Esmaeilpour, Z., Malmir, N., Mahdavifar-Khayati, R., Soleimani, G., Cha, Y.-H., Shereen, A.D., Nitsche, M.A., *et al.*: fmri and transcranial electrical stimulation (tes): A systematic review of parameter space and outcomes. Progress in Neuro-Psychopharmacology and Biological Psychiatry **107**, 110149 (2021)

[27] Zheng, W., Bao, C., Mu, R., Wang, J., Li, T., Zhao, Z., Yao, Z., Hu, B.: Frequency-specific dual-attention based adversarial network for blood oxygen level-dependent time series prediction. Human Brain Mapping **45**(14), 70032 (2024)

[28] Sun, Y., Cabezas, M., Lee, J., Wang, C., Zhang, W., Calamante, F., Lv, J.: Predicting human brain states with transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 136–146 (2024). Springer

[29] Hu, Y., Jiang, Y., Li, W., Yuan, Y.: Synthesizing realistic fmri: a physiological dynamics-driven hierarchical diffusion model for efficient fmri acquisition. In: The Thirteenth International Conference on Learning Representations (2025)

[30] Sun, Y., Chahine, D., Wen, Q., Liu, T., Li, X., Yuan, Y., Calamante, F., Lv, J.: Voxel-level brain states prediction using swin transformer. arXiv preprint arXiv:2506.11455 (2025)

[31] Müller, R.-A., Shih, P., Keehn, B., Deyoe, J.R., Leyden, K.M., Shukla, D.K.: Underconnected, but how? a survey of functional connectivity mri studies in autism spectrum disorders. Cerebral cortex **21**(10), 2233–2243 (2011)

[32] Elliott, M.L., Romer, A., Knodt, A.R., Hariri, A.R.: A connectome-wide functional signature of transdiagnostic risk for mental illness. Biological psychiatry **84**(6), 452–459 (2018)

[33] Sun, L., Zhao, T., Liang, X., Xia, M., Li, Q., Liao, X., Gong, G., Wang, Q., Pang, C., Yu, Q., et al.: Functional connectome through the human life span. BioRxiv, 2023–09 (2024)

[34] Sun, L., Zhao, T., Liang, X., Xia, M., Li, Q., Liao, X., Gong, G., Wang, Q., Pang, C., Yu, Q., et al.: Human lifespan changes in the brain's functional connectome. Nature neuroscience, 1–11 (2025)

[35] Luo, N., Sui, J., Abrol, A., Lin, D., Chen, J., Vergara, V.M., Fu, Z., Du, Y., Damaraju, E., Xu, Y., *et al.*: Age-related structural and functional variations in 5,967 individuals across the adult lifespan. Human brain mapping **41**(7), 1725–1737 (2020)

[36] Hassanzadeh, E., Ailion, A., Hassanzadeh, M., Hornak, A., Peled, N., Martino, D., Warfield, S.K., Lan, Z., Gholipour, T., Stufflebeam, S.M.: Imaging and anesthesia protocol optimization in sedated clinical resting-state fmri. American Journal of Neuroradiology **46**(2), 293–301 (2025)

[37] Seeliger, K., Sommers, R., Güçlü, U., Bosch, S., Van Gerven, M.: A large single-participant fmri dataset for probing brain responses to naturalistic stimuli in space and time. BioRxiv, 687681 (2019)

[38] Fox, M.D., Raichle, M.E.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature reviews neuroscience **8**(9), 700–711 (2007)

[39] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

[40] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.*: The uk biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)

[41] Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., *et al.*: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)

[42] Somerville, L.H., Bookheimer, S.Y., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Dapretto, M., Elam, J.S., Gaffrey, M.S., Harms, M.P., Hodge, C., *et al.*: The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. Neuroimage **183**, 456–468 (2018)

[43] Bookheimer, S.Y., Salat, D.H., Terpstra, M., Ances, B.M., Barch, D.M., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Diaz-Santos, M., Elam, J.S., *et al.*: The lifespan human connectome project in aging: an overview. Neuroimage **185**, 335–348 (2019)

[44] Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S.: A whole brain fmri atlas generated via spatially constrained spectral clustering. Human brain mapping **33**(8), 1914–1928 (2012)

[45] Cai, W., Liang, Y., Liu, X., Feng, J., Wu, Y.: Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 11141–11149 (2024)

[46] Wang, Y., Xu, Y., Yang, J., Wu, M., Li, X., Xie, L., Chen, Z.: Fully-connected spatial-temporal graph for multivariate time-series data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 15715–15724 (2024)

[47] Yi, K., Zhang, Q., Fan, W., He, H., Hu, L., Wang, P., An, N., Cao, L., Niu, Z.: Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. Advances in neural information processing systems **36**, 69638–69660 (2023)

[48] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625 (2023)

[49] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. arXiv preprint arXiv:2211.14730 (2022)

[50] Chen, S.-A., Li, C.-L., Yoder, N., Arik, S.O., Pfister, T.: Tsmixer: An all-mlp architecture for time series forecasting. arXiv preprint arXiv:2303.06053 (2023)

[51] Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 11121–11128 (2023)

[52] Zhang, T., Zhang, Y., Cao, W., Bian, J., Yi, X., Zheng, S., Li, J.: Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. arxiv. arXiv preprint arXiv:2207.01186 (2022)

[53] Dong, J., Wu, H., Zhang, H., Zhang, L., Wang, J., Long, M.: Simmtm: A simple pre-training framework for masked time-series modeling. Advances in Neural Information Processing Systems **36**, 29996–30025 (2023)

[54] Zhou, T., Niu, P., Sun, L., Jin, R., *et al.*: One fits all: Power general time series analysis by pretrained lm. Advances in neural information processing systems **36**, 43322–43355 (2023)

[55] Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., et al.: The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of neurophysiology (2011)

[56] Tong, W., Li, Y.-X., Zhao, X.-Y., Chen, Q.-Q., Gao, Y.-B., Li, P., Wu, E.Q.: fmri-based brain disease diagnosis: A graph network approach. IEEE Transactions on Medical Robotics and Bionics **5**(2), 312–322 (2023)

[57] Liu, M., Zhang, H., Liu, M., Chen, D., Zhuang, Z., Wang, X., Zhang, L., Peng, D., Wang, Q.: Randomizing human brain function representation for brain disease diagnosis. IEEE Transactions on Medical Imaging (2024)

[58] Huang, Z.-A., Zhu, Z., Yau, C.H., Tan, K.C.: Identifying autism spectrum disorder from resting-state fmri using deep belief network. IEEE Transactions on neural networks and learning systems **32**(7), 2847–2861 (2020)

[59] Rosenblatt, M., Tejavibulya, L., Sun, H., Camp, C.C., Khaitova, M., Adkinson, B.D., Jiang, R., Westwater, M.L., Noble, S., Scheinost, D.: Power and reproducibility in the external validation of brain-phenotype predictions. Nature Human Behaviour **8**(10), 2018–2033 (2024)

[60] Chandrashekar, P.B., Alatkar, S., Wang, J., Hoffman, G.E., He, C., Jin, T., Khullar, S., Bendl, J., Fullard, J.F., Roussos, P., *et al.*: Deepgami: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction. Genome Medicine **15**(1), 88 (2023)

[61] He, Z., Li, W., Jiang, Y., Peng, Z., Wang, P., Li, X., Liu, T., Han, J., Zhang, T., Yuan, Y.:

F2tnet: Fmri to t1w mri knowledge transfer network for brain multi-phenotype prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 265–275 (2024). Springer

[62] Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A.A.C., Lukemire, J., Zhan, L., He, L., Guo, Y., Yang, C.: Braingb: a benchmark for brain network analysis with graph neural networks. IEEE transactions on medical imaging **42**(2), 493–506 (2022)

[63] Yang, Y., Cui, H., Yang, C.: Ptgb: Pre-train graph neural networks for brain network analysis. arXiv preprint arXiv:2305.14376 (2023)

[64] Su, W.-C., Culotta, M., Mueller, J., Tsuzuki, D., Bhat, A.N.: Autism-related differences in cortical activation when observing, producing, and imitating communicative gestures: An fnirs study. Brain Sciences **13**(9), 1284 (2023)

[65] Benkarim, O., Paquola, C., Park, B.-y., Hong, S.-J., Royer, J., Wael, R., Lariviere, S., Valk, S., Bzdok, D., Mottron, L., *et al.*: Connectivity alterations in autism reflect functional idiosyncrasy. Communications biology **4**(1), 1078 (2021)

[66] Steinbach, M., Tan, P.-N.: kNN: K-nearest Neighbors, pp. 165–176. Chapman and Hall/CRC, Boca Raton (2009)

[67] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016)

[68] Kipf, T.N., Welling, M.: Variational graph auto-encoders (2016)

[69] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)

[70] Leon, M.: Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. Information Systems, 102620 (2025)

[71] Sozer, A., Sahin, M.C., Sozer, B., Erol, G., Tufek, O.Y., Nernekli, K., Demirtas, Z., Celtikci, E.: Do llms have 'the eye'for mri? evaluating gpt-4o, grok, and gemini on brain mri performance: First evaluation of grok in medical imaging and a comparative analysis. Diagnostics **15**(11), 1320 (2025)

[72] Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., et al.: Advancing multimodal medical capabilities of gemini. arXiv preprint arXiv:2405.03162 (2024)

[73] Niu, W., Xie, Z., Sun, Y., He, W., Xu, M., Hao, C.: Langtime: A language-guided unified model for time series forecasting with proximal policy optimization. arXiv preprint arXiv:2503.08271 (2025)

[74] Chen, Y., Gandhi, R., Zhang, Y., Fan, C.: Nl2tl: Transforming natural languages to temporal logics using large language models. arXiv preprint arXiv:2305.07766 (2023)

[75] Liu, C., Xu, Q., Miao, H., Yang, S., Zhang, L., Long, C., Li, Z., Zhao, R.: Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 18780–18788 (2025)

[76] Mizuno, T., Fujimoto, S., Ishikawa, A.: Generation of individual daily trajectories by gpt-2. Frontiers in Physics **10**, 1021176 (2022)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf